О нижней границе вероятности ошибки классификации данных от ансамбля источников

М.М. Ланге[1], С.В. Парамонов[1]

[1]*ФИЦ "Информатика и управление" РАН, Вавилова 40, Москва, Россия, 119333*

Аннотация

Исследуется модель классификации объектов от ансамбля источников в терминах зависимости вероятности ошибки от количества информации, содержащейся в множестве классифицируемых объектов. Приводится аналитическая нижняя граница средней вероятности ошибки как функция средней взаимной информации между ансамблем источников и множеством решений о классах объектов. Для любого набора разделяющих функций вводится избыточность средней вероятности ошибки относительно нижней границы. Демонстрируется возможность уменьшения средней вероятности ошибки и избыточности за счет использования различных композиций разделяющих функций на ансамбле изображений лиц и подписей.

Ключевые слова

классификация, ансамбль источников, вероятность ошибки, взаимная информация, разделяющие функции, избыточность, изображение

1. Введение

Одна из фундаментальных проблем классификации данных состоит в получении нижней границы вероятности ошибки по составным объектам от ансамбля источников различной модальности. В работе [1] приведены оценки достижимой точности классификации для различных решающих алгоритмов, но не проведено сравнение этих оценок с потенциально достижимой точностью классификации для заданного источника данных. Нижняя граница вероятности ошибки как функция количества информации, содержащейся в множестве объектов от заданного источника, предложена в работе [2]. Эта граница базируется на модели кодирования источника с допустимой погрешностью при наличии канала наблюдения с искажениями [3] и является обобщением границы Шеннона для скорости кодирования конечного множества дискретных сообщений с хемминговой мерой погрешности [4].

В настоящей работе дается обобщение полученной в [2] нижней границы для ансамбля источников и предлагается методика оценивания избыточности средней вероятности ошибки относительно нижней границы для различных композиций разделяющих функций.

2. Основной результат

Рассматривается схема $\Omega \to \mathbf{X}^M \to \hat{\Omega}$, в которой Ω и $\hat{\Omega}$ – множества классов и решений о классах для составных объектов $\mathbf{x}^M = (\mathbf{x}_1, ..., \mathbf{x}_M)$ от $M \geq 1$ источников (по одному объекту одного класса от каждого источника), образующих ансамбль $\mathbf{X}^M = \mathbf{X}_1 ... \mathbf{X}_M$. Получено соотношение между наименьшей средней взаимной информацией $I(\mathbf{X}^M; \hat{\Omega}) \geq 0$ и средней вероятностью ошибки $E(\mathbf{X}^M, \hat{\Omega}) \leq \varepsilon$ в форме убывающей функции

$$R_M(\varepsilon) = I(\mathbf{X}^M; \Omega) - h(\varepsilon - \varepsilon^M_{\min}) - (\varepsilon - \varepsilon^M_{\min}) \ln(c-1), \; \varepsilon^M_{\min} \leq \varepsilon \leq \varepsilon^M_{\max}, \quad (1)$$

где $h(z) = -z \ln z - (1-z)\ln(1-z)$, $R_M(\varepsilon^M_{\min}) = I(\mathbf{X}^M; \Omega)$, $R_M(\varepsilon^M_{\max}) = 0$ и $I(\mathbf{X}^M; \Omega)$ – средняя взаимная информация между $\mathbf{X}^M$ и Ω. Обращение функции (1) дает нижнюю границу средней вероятности ошибки $\varepsilon = R_M^{-1}(I(\mathbf{X}^M; \hat{\Omega}))$ при значениях $I(\mathbf{X}^M; \hat{\Omega}) \leq I(\mathbf{X}^M; \Omega)$. Функция (1)

является обобщением границы Шеннона [4], когда $I(\mathbf{X}^M;\Omega)=H(\Omega)$ и $\varepsilon_{\min}^M=0$, где $H(\Omega)\le\ln c$ – энтропия множества Ω мощности $c\ge 2$. Характеристики источников и ансамбля удовлетворяют неравенствам: $\max_{m=1}^M I(\mathbf{X}_m;\Omega)<I(\mathbf{X}^M;\Omega)$ и $\min_{m=1}^c \varepsilon_{\min\_m}>\varepsilon_{\min}^M$. В случае равновероятных классов, для любого источника и ансамбля имеем $\varepsilon_{\max\_m}=\varepsilon_{\max}^M=(c-1)/c$.

Обращение границы (1) использовано для оценивания избыточности средней вероятности ошибки классификации на ансамбле $\mathbf{X}^M$ по заданному набору разделяющих функций $\mathrm{G}(\mathbf{X}^M)=\left\{g_j(\mathbf{x}^M),\mathbf{x}^M\in\mathbf{X}^M\right\}_{j=1}^c$. Набору $\mathrm{G}(\mathbf{X}^M)$ соответствуют средняя взаимная информация $I_{\mathrm{G}}(\mathbf{X}^M;\hat{\Omega})$ и средняя вероятность ошибки $E_{\mathrm{G}}(\mathbf{X}^M,\hat{\Omega})$, которые дают избыточность

$$r_{\mathrm{G}}=E_{\mathrm{G}}(\mathbf{X}^M,\hat{\Omega})-R_M^{-1}(I_{\mathrm{G}}(\mathbf{X}^M;\hat{\Omega})).$$

(2)

Разделяющие функции на ансамбле задаются в форме композиций разделяющих функций для источников так, что $g_j(\mathbf{x}^M)=\prod_{m=1}^M\prod_{k=1}^{K_m}g_{jk}(\mathbf{x}_m)$. Здесь $\mathrm{G}_k(\mathbf{X}_m)=\left\{g_{jk}(\mathbf{x}_m),\mathbf{x}_m\in\mathbf{X}_m\right\}_{j=1}^c$ – k-й слабый набор разделяющих функций на множестве $\mathbf{X}_m$, для которого $I_{\mathrm{G}_k}(\mathbf{X}_m;\hat{\Omega})\ll I(\mathbf{X}_m;\Omega)$; K_m слабых наборов на множестве $\mathbf{X}_m$ порождают композитный набор $\mathrm{G}^{K_m}(\mathbf{X}_m)$ мощности K_m.

Используя евклидово расстояние объектов, заданных изображениями, построены условные по классам распределения вероятностей на множествах изображений лиц и подписей. Для равновероятных классов и условных распределений объектов, получены численные реализации границы вида (1) для указанных источников и ансамбля. Вычислены характеристики $I_{\mathrm{G}}(\mathbf{X}^M;\hat{\Omega})$, $E_{\mathrm{G}}(\mathbf{X}^M,\hat{\Omega})$ и соответствующие значения избыточности (2) для различных разделяющих функций. Полученные численные оценки показали увеличение $I_{\mathrm{G}}(\mathbf{X}^M;\hat{\Omega})$ и уменьшение $E_{\mathrm{G}}(\mathbf{X}^M,\hat{\Omega})$ и r_{G} на ансамбле источников с ростом мощности композитных наборов разделяющих функций на множествах изображений лиц и подписей.

3. Заключение

В рамках теоретико-информационной модели классификации составных объектов от ансамбля источников предложена нижняя граница средней вероятности ошибки как функция средней взаимной информации между ансамблем и множеством решений о классах предъявляемых объектов. Полученная граница является аналогом нижней границы для известной в теории информации функции скорость-погрешность (Rate Distortion Function) при наличии стохастического мульти канала наблюдения, который связывает множество классов и множество объектов от ансамбля источников. Независимость построенной границы от решающего правила позволяет оценивать эффективность решающих алгоритмов с заданными разделяющими функциями в терминах избыточности средней вероятности ошибки относительно нижней границы. Избыточность демонстрирует отклонение вероятности ошибки от потенциально возможного значения при количестве информации, реализуемом разделяющими функциями. Продемонстрирована возможность уменьшения вероятности ошибки и избыточности за счет применения композиций различных разделяющих функций. Предложенный подход допускает развитие с использованием многоуровневых разделяющих функций, которые могут найти применение в решающих деревьях и нейронных сетях.

4. Литература

[1] Kuncheva, L.I. Limits on the majority vote accuracy in classifier fusion / L.I. Kuncheva, C.J. Whitaker, C.A. Shipp, R.P.W Duin // Pattern Analysis and Applications. – 2003. – Vol. 6. – P. 22-31. DOI: 10.1007/s10044-002-0173-7.

[2] Lange, M. On Data Classification Efficiency Based a Trade-off Relation between Mutual Information and Error Probability / A. Lange, S. Paramonov // International Conference on

Information Technology and Nanotechnology (ITNT). IEEE Proceedings. – 2020. DOI: 10.1109/ITNT49337.2020.9253225.

[3] Dobrushin, R.L. Information transmission with additional noise / R.L. Dobrushin, B.S. Tsybakov // IRE Trans. Information Theory. – 1962. – Vol. 8(5). – P. 293-304. DOI: 10.1109/ TIT.1962.1057738.

[4] Gallager, R.G. Information Theory and Reliable Communication. – New York: Wiley & Sons, Inc., 1968. – 588 p.